

PROJECT PRESENTATION

# INSTACART MARKET BASKET ANALYSIS



# Business Problem

## SETTING

A grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them

## CURRENT METHODS

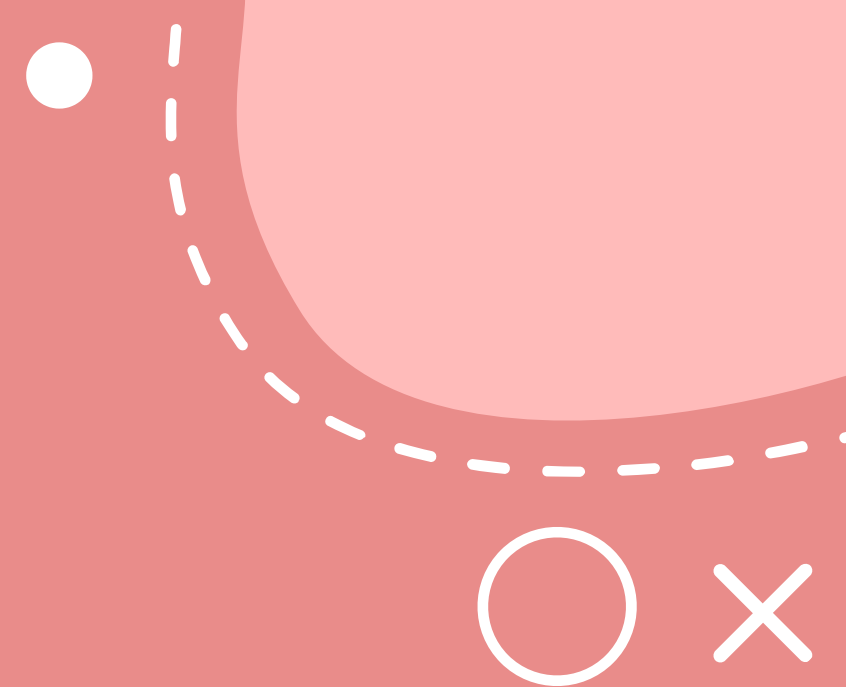
Currently they use transactional data to develop models that predict which products a user will buy again, try for the first time, or add to their cart next during a session

## RESEARCH QUESTIONS ADDRESSED

Predict whether a product will be reordered or not in the future by the customer  
Predict which department will the next product ordered belong to

## SOLUTION

The ability to identify which products the customers are likely to purchase again, and automatically adding those to cart through obtained predictions or provide a seamless interface for doing so will enhance their user experience



# DATA EXPLORATION

MISSING VALUE IMPUTATION  
OUTLIER TREATMENT  
DISTRIBUTION ANALYSIS  
CORRELATION TESTS  
DATA VISUALIZATION  
DATA PREPARATION

## ORDER RELATED FEATURES

ORDER\_ID

ORDER\_NUMBER

AVERAGE\_DAYS\_BETWEEN\_ORDERS

NB\_ORDERS(NUMBER OF ORDERS)

## TIME RELATED FEATURES

ORDER - HOUR OF THE DAY

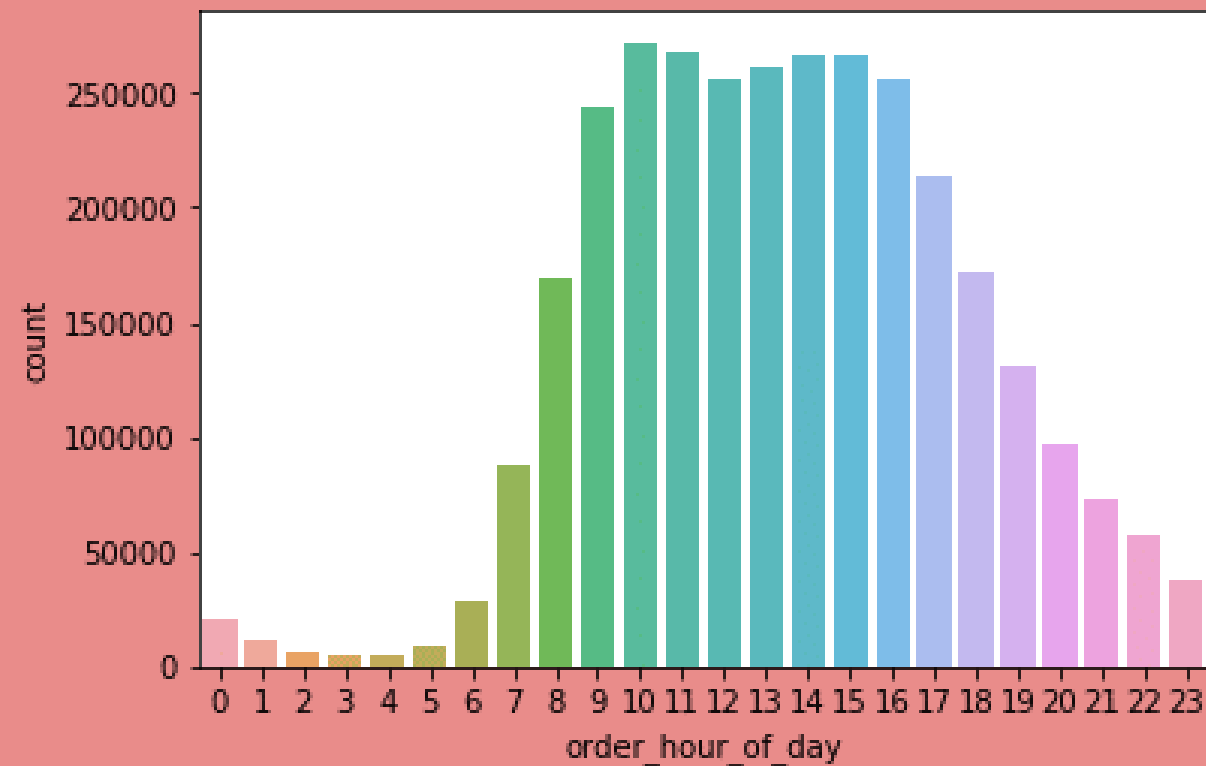
ORDER - DAY OF THE WEEK

DAYS SINCE PRIOR ORDER

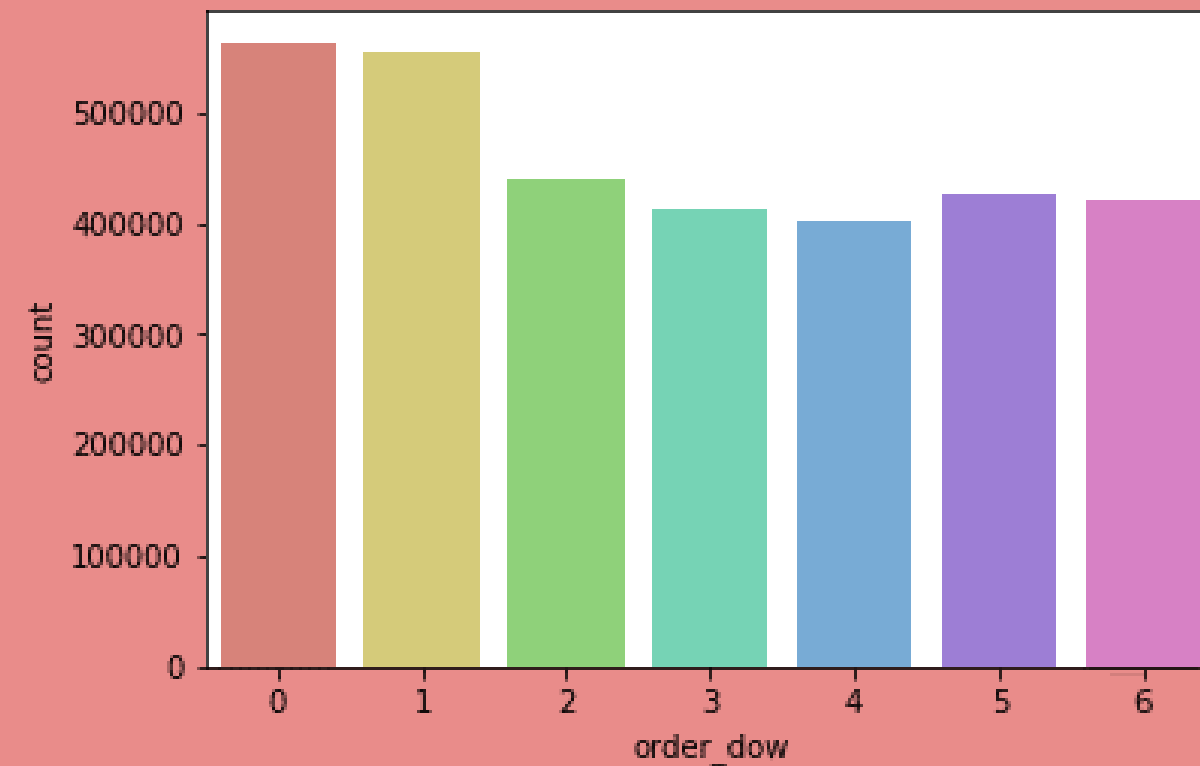
DAYS SINCE RATIO



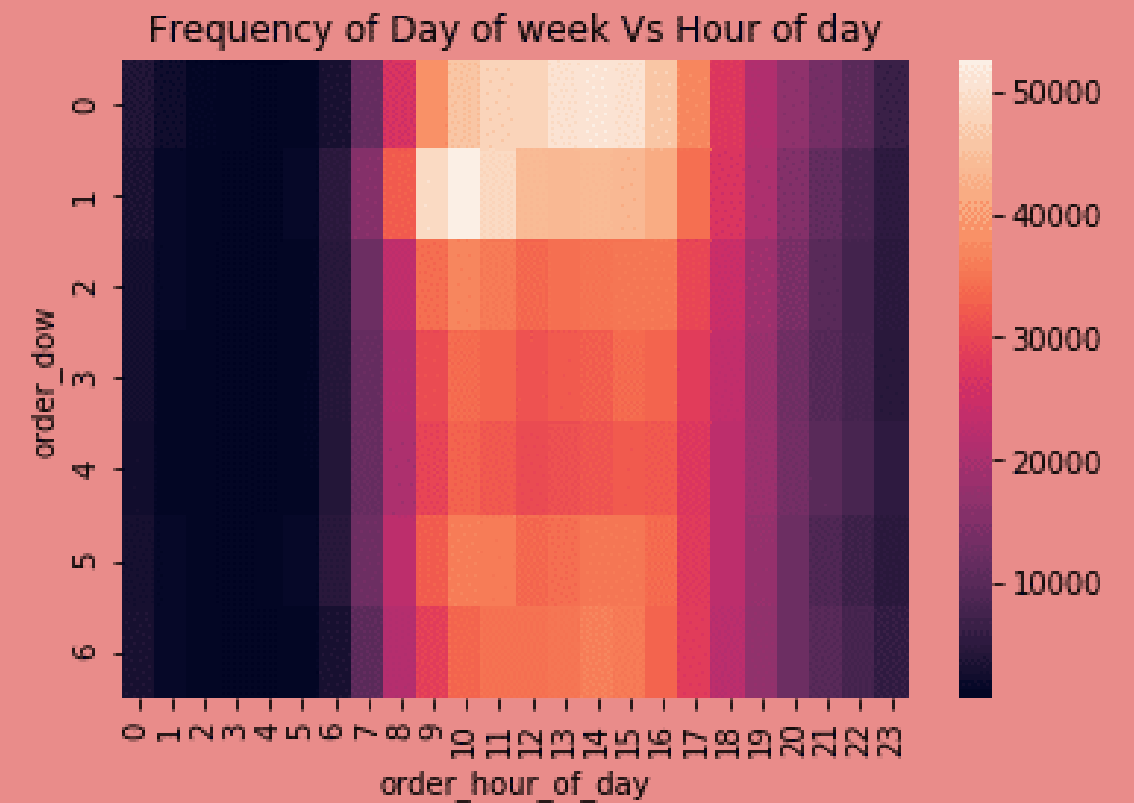
# DATA FOR THE HUMAN EYE



HOUR OF THE DAY

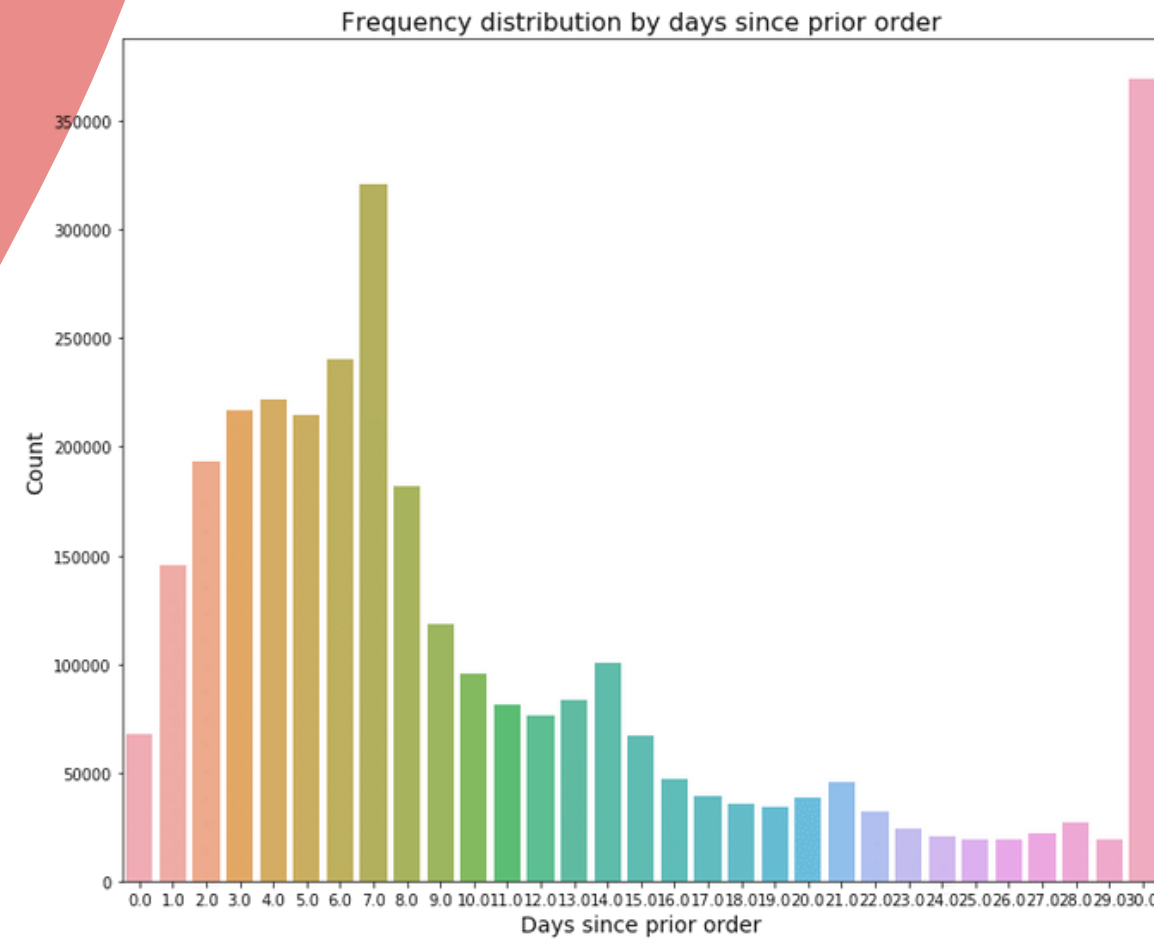


DAY OF WEEK

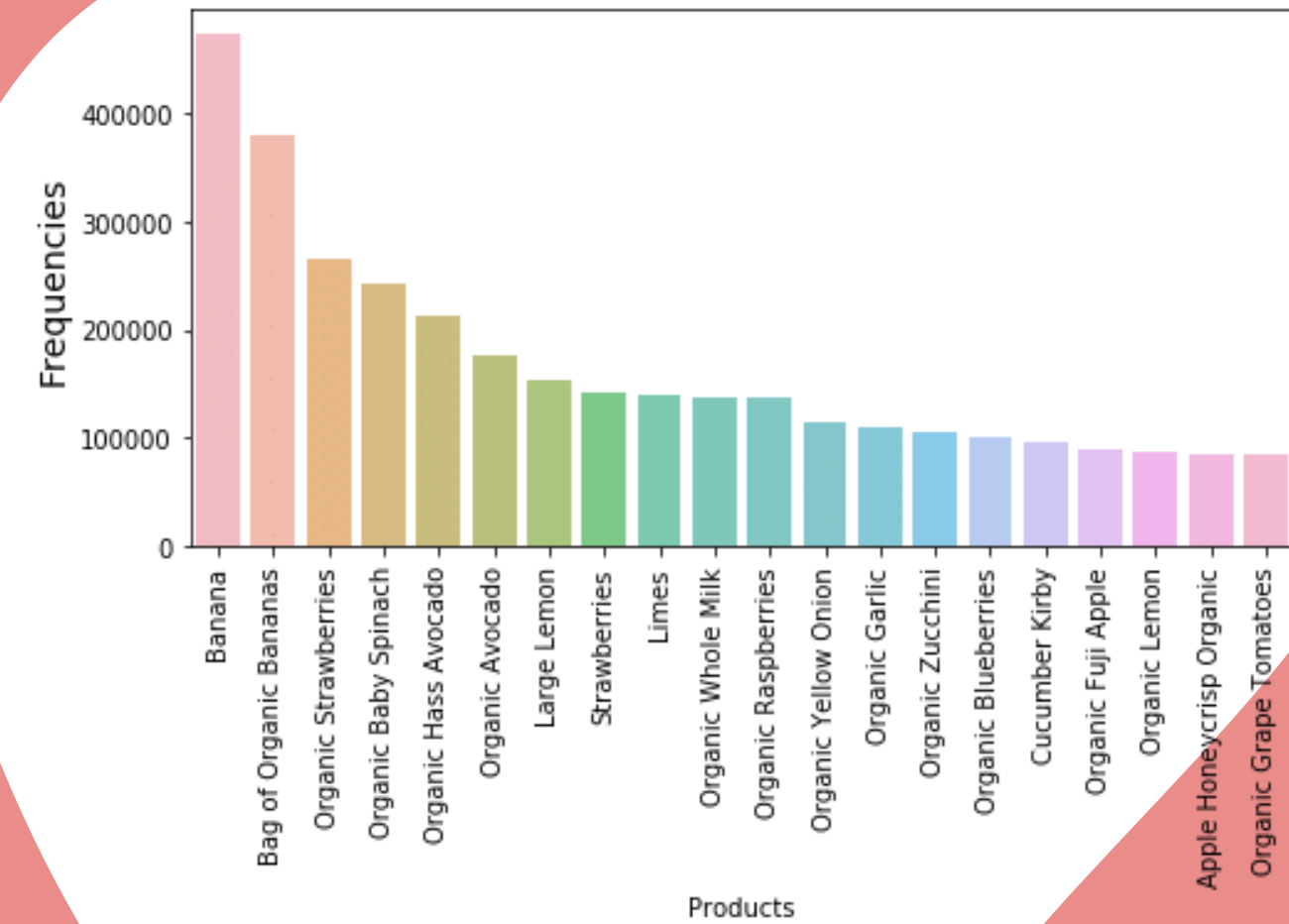


Day of week  
vs  
Hour of the day

# DEEPER DIVE



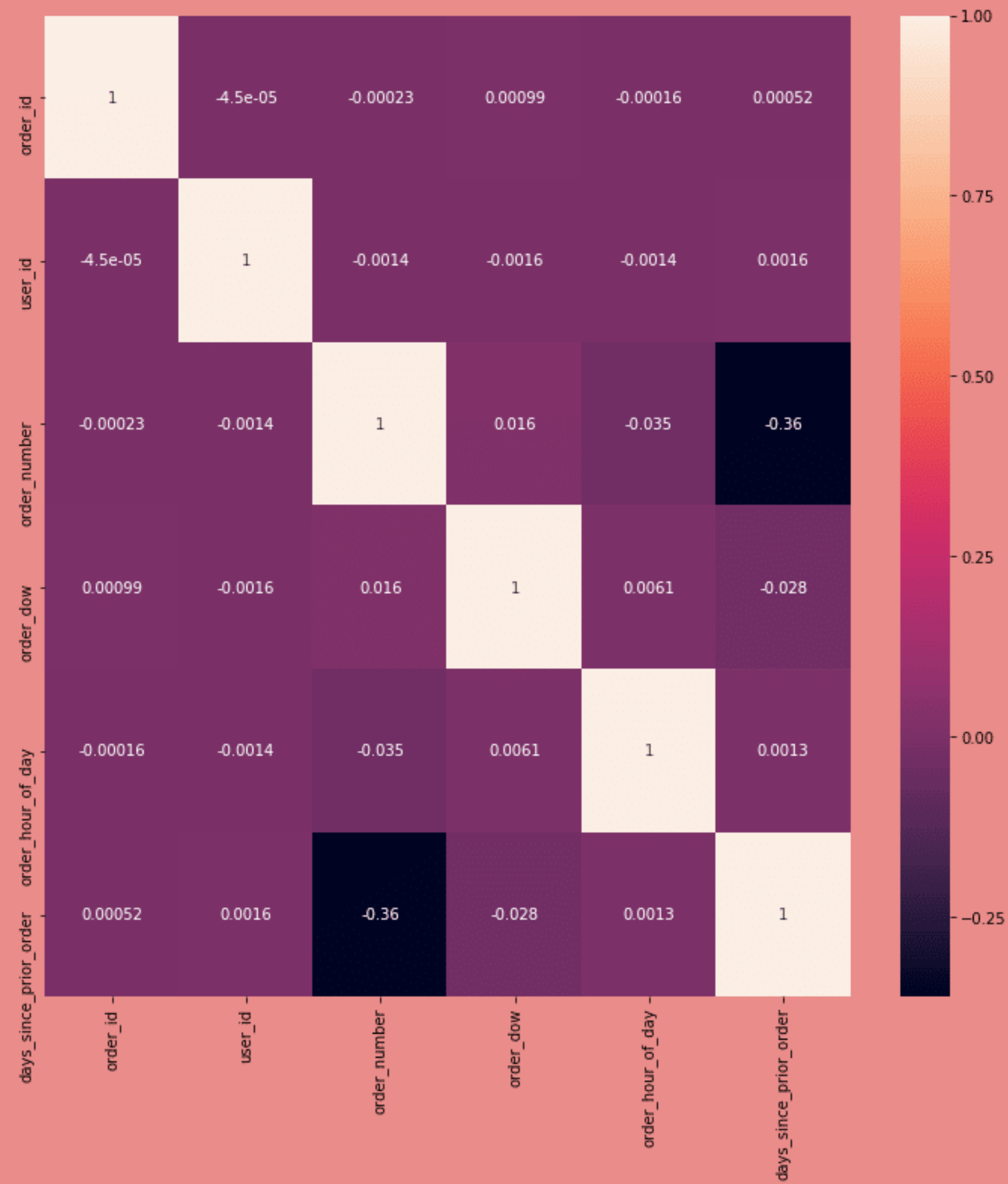
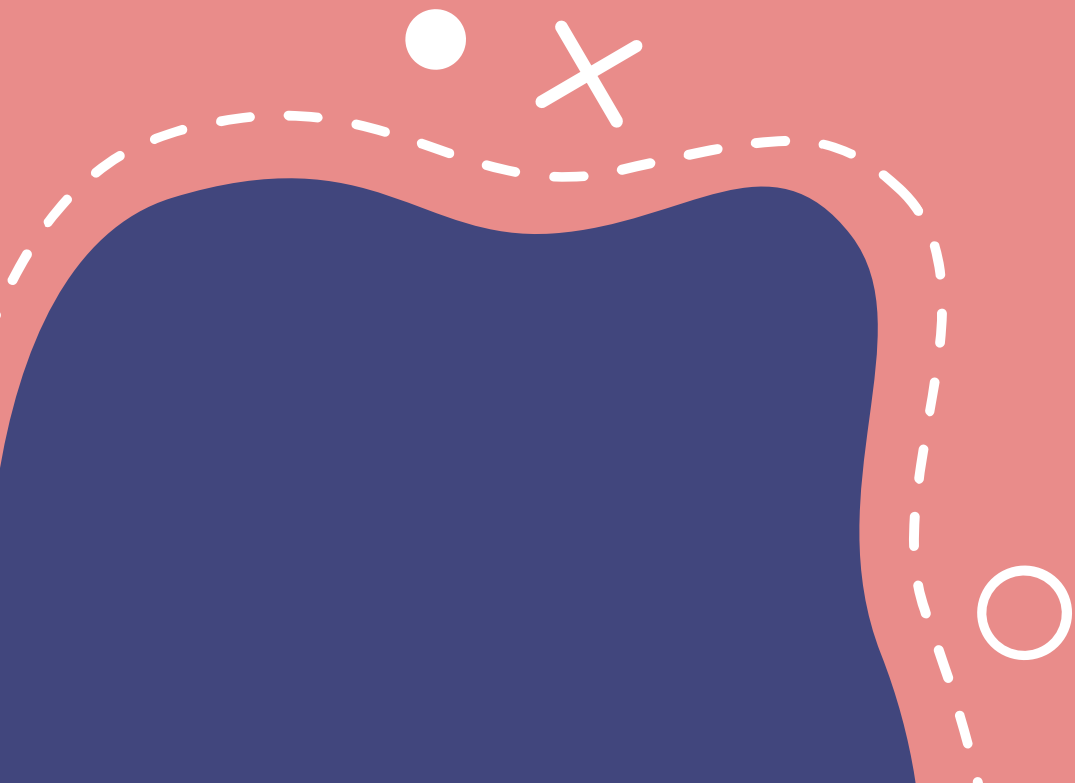
D. Since Prior Order



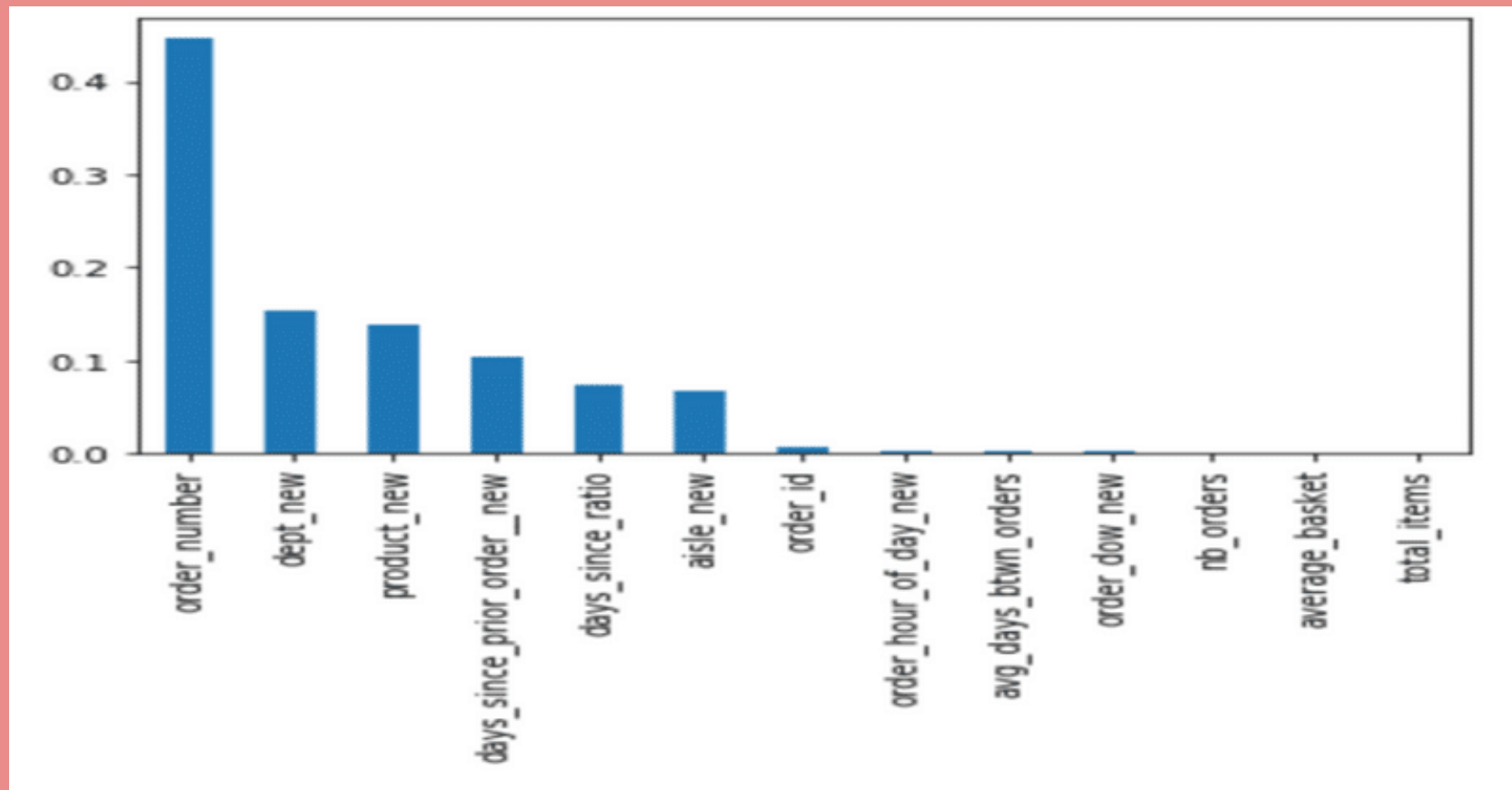
Top 20 products

More in the report...

# CORRELATION STUDY



# VARIABLE IMPORTANCE



ORDER NUMBER

DEPARTMENT

PRODUCT

DAYS SINCE PRIOR  
ORDER

DAYS SINCE RATIO

AISLE

ORDER ID

ORDER HOUR OF THE DAY

AVG DAYS B/W ORDERS





# MODEL BUILDING

## CHOICES

Presentations are tools that can be used as lectures. We used various models to predict if a product will be reordered or not, such as, Logistic regression, Random Forest, Adaboost Classifier and Gradient Boosting. Logistic regression was the worst performing model while Gradient Boosting was the best performing model with an Accuracy of 67.07% on the test data. It also had the highest AUC of 0.66. Listed below are the metrics for each model on the training as well as the test data

MODEL	ACC.	PRECISION	RECALL	AUC
Logistic Regression	59.70%	85.63%	61.83%	0.56
Random Forest	59.76%	85.63%	61.83%	0.56
Adaboost Classifier	65.57%	80.02%	68.06%	0.65
Gradient Boosting	67.55%	82.87% <sub>s</sub>	69.08%	0.66



# QUESTIONS #?!

NO PRESSURE THOUGH

